Translational Cancer Research Data Quality

– The Context Factor

A DISSERTATION SUBMITTED TO

THE FACULTY OF UNIVERSITY OF MINNESOTA

BY

Giordi Orreggio

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

STUART M. SPEEDIE, PH.D., FACMI (Adviser)

August 2017

ACKNOWLEDGEMENTS

Many people have been influential in the successful completion of this project: Dr. Stuart Speedie, my faculty advisor, for his tireless guidance and support. Without his mentoring and patience over many years of this work, this project could not have been completed. I would also like to thank the long-standing members of the preliminary oral and final examination committees, including Dr. Sarah Cooley, for her feedback, support and advice through this journey.

DEDICATION

To my mother, Delroy Louise Dixon-O'Reggio, for her decades of belief in me and her selfless love, up until her final battle with pancreatic cancer, April 9, 2017. To my husband, Nicholas Keith Fradgely, who is anxiously awaiting to announce to our family in Britain, "My husband the doctor!" To our children, Emma, Mary Ellen, and Luke—a generation of technology and hope.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

While data may be objectively "complete" and conforming regarding lists of known valid values, failure to check for contextual inconsistencies potentially result in data that is less efficient at leading to knowledge than it may appear, especially when data sets from different sources are integrated for secondary use. To examine this problem and our proposed solution, we split our research into three distinct parts, which will be introduced in the three chapters throughout this paper.

In our first of the related studies, we introduce a novel yet simple and fast method of increasing the quality of translational cancer research data. The method is novel in that cross checking we label as a contextual method is used to measure data quality. We compare the results to those of traditional methods of checking accuracy and completeness, something we label as an intrinsic method.

In our second study, we introduce an algorithm and method to computationally identify contextual relationships within a dataset. We compare the results to those using subject matter experts (SMEs), a gold standard in data quality efforts.

Finally, we end this group of research with our third study. In this case study, we apply our methods to two datasets, each meant to convey the same information but coming from different sources. Then, we further test our methods by applying

them to a third dataset, the combination of the first two datasets, to see how well our methods identify the data quality of integrated data sets.

Combined this set of studies investigates a novel method, involvement of SMEs, and data integration regarding translational cancer research data quality. Our perspective is a focus on “translational cancer research data quality" due to each of its component parts’ overall relevancy to the subject.

We focus on *transitional research* data quality because, as opposed to common knowledge, the knowledge flowing from research is knowledge that is in flux, evolving, and novel. Translational medicine is a rapidly growing discipline in biomedical research and aims to expedite the discovery of new diagnostic tools and treatments by using a multi-disciplinary, highly collaborative, "bench-to-bedside" approach [1]. The length of time this new data has been available for data quality testing, therefore, is less than that of established data.

Furthermore, as a national medical priority, *cancer research* is relatively well-funded [2]. With cancer research, so fueled, the data volume increase in this domain is high when compared to other data areas.

We focus on *data*, of course, because the volume of data in general is increasing as computational power and speed increase [3]. Concurrently, storage capability is becoming both more efficient and more affordable.

We focus on data *quality*. Tradition focuses on accuracy and completeness. However, with increased volume there is increased opportunity to acquiring confidence in the quality of a data set through cross-checking. Cronbach's alpha is a measure of internal consistency, that is, how closely related a set of items are as a group. Cronbach's alpha can be written as a function of the number of test items and the average inter-correlation among the items. Below, for conceptual purposes, we show the formula for the standardized Cronbach's alpha:

$$\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}$$

Here N is equal to the number of items, c-bar is the average inter-item covariance among the items and v-bar equals the average variance. One can see from this formula that if you increase the number of items, you increase Cronbach's alpha.

ACRONYMS AND FRAMEWORK

Acronyms referenced in this group of studies fall into one of two domains: oncology and health informatics. Acronyms within the oncology domain include those for the terms: absolute lymphocyte (ALYM), natural killer [cell] (NK), absolute NK (ANK), complete blood count (CBC), Masonic Cancer Center (MCC), National Cancer Institute (NCI), Oncology Business Layer (OBL), Oncology Medical Informatics and Services

(OMIS), and white blood cell (WBC). Acronyms within the health informatics domains include those for the terms: Common Data Elements (CDE), conditional functional dependency (CFD), electronic health record (EHR), information technology (IT), data quality (DQ), contextual DQ (CDQ), and intrinsic DQ (IDQ).

We develop, describe, and apply two frameworks – the IDQ Framework and the CDQ Framework to expose and quantify DQ issues in three case studies. Our IDQ Framework represents a traditional method to detecting DQ issues. Categories of measures within this method include: missing data, i.e. incompleteness, and intrinsic inaccuracies. Our CDQ Framework represents our proposed novel augmentation to traditional methods. We compare the results of traditional DQ issue detection efforts augmented with our CDF method versus those from traditional methods alone. Ultimately, our studies' goal is to show that by utilizing the combined methodology, the quality of data, and as such the value, can be improved.

UNIFIED LITERATURE REVIEW

Second only to heart disease, cancer is the leading cause of death worldwide [4]. In Taiwan, where cancer has become the leading cause of death since 1982, maintaining a high-quality cancer registry database is essential, and efforts have ensured completeness of this data at the level of 97% [5]. This focus on an intrinsic, non-contextual measure – accuracy – is not unusual. Electronic validation of oncology clinical trial data is typically performed on isolated data elements at the transaction level, prior to data integration, around the time of data entry into a system, whether input is manual or automated (e.g. as in a transmission, or a health level 7 message). Health data quality (DQ) studies have investigated diagnostic coding [6] and, other health data [7] [8]. DQ improvement efforts tend to focus narrowly on accuracy [9] [10] [11] and completeness [12] [13]. Accuracy-related measures such as electronic sources matching electronic targets, precision, value within range, and allowed values, and completeness-related measures such as lack of missing values and completed required fields are typical measures of electronic DQ [12] [13]. These measures are intrinsic in nature since they are characteristic of a particular data item. A potential shortcoming with this approach is a lack of contextual sensitivity.

The need for contextual sensitivity becomes greater for translational cancer research data considering the amount of data integration involved. Researchers at the University of Arkansas, for example, understand that because clinical research data is the output of a federation of collection mechanisms and systems, there is an increased risk of poor data quality leading to inefficient use of research data, or the need for costly repetition of clinical studies [14]. The researchers present two tools for improving data quality of clinical research data relying on the National Cancer Institute's Common Data Elements (CDE) as a standard representation of possible questions and data elements to A: automatically suggest CDE annotations for already collected data based on semantic and syntactic analysis utilizing the Unified Medical Language System (UMLS) Terminology Services' Metathesaurus and B: annotate and constrain new clinical research questions though a simple-to-use "CDE Browser." The results showed that a small portion of suggested annotations were syntactically and semantically sound; however, many of the results were complete misses. Lacking in the researchers' approach was a way to establish sensitivity to context.

Our three studies build on past efforts to provide automated data quality assessment, improvement, and constraints for clinical research data, by first confirming the importance of context sensitivity to data quality efforts, next by exploring a way to automate context sensitivity, and finally by testing the application of context-sensitive data quality assessment to federated data.

CHAPTER II: CONTEXTUAL VS. INTRINSIC EVALUATION OF CANCER RESEARCH DQ

Giordi Orreggio, MHI[1], Sarah Cooley, MD[1], Stuart Speedie, PhD[2]

[1]Masonic Cancer Center, University of Minnesota, Minneapolis, MN;

[2]Institute of Health Informatics, Minneapolis, MN

SYNOPSIS

Traditional data quality (DQ) efforts focus on intrinsic measures, such as accuracy and completeness. Translational research in oncology relies heavily on integration of routine patient care data with research laboratory generated data. The repurposing of such data for research use raises the risk of introducing contextual DQ (CDQ) issues. CDQ issues include logical inconsistency and improbable distributions of data element values. Methods were developed to assess the CDQ of 6442 absolute natural killer cell (ANK) blood sample collection records. This paper highlights this specific example to present a novel method of exposing DQ issues. Compared to traditional intrinsic tests of DQ, which exposed problems in 1161 (18%) of the records, CDQ testing exposed an additional 3177 records, or 4338 (67%) records with some concern about data quality.

INTRODUCTION

Electronic validation of oncology clinical trial data is typically performed on isolated data elements at the transaction level, prior to data integration, around the time of data entry into a system, whether input is manual or automated (e.g. as in a transmission, or a health level 7 message). Health data quality (DQ) studies have investigated diagnostic coding [6] and, other health data [7] [8]. DQ improvement efforts tend to focus narrowly on accuracy [9] [10] [11] and completeness [12] [13]. Accuracy-related measures such as electronic sources matching electronic targets, precision, value within range, and allowed values, and completeness-related measures such as lack of missing values and completed required fields are typical measures of electronic DQ [12] [13]. These measures are intrinsic in nature since they are characteristic of a particular data item.

One cannot assume that if the data appears to be intrinsically accurate and complete at time of entry, the quality of repurposed and integrated data will be optimal. The problem regarding translational oncology research is that because this domain relies heavily on integration of repurposed routine patient care data with research data, intrinsic methods fall short. As data collected for patient care is reused for a different purpose (research), this raises the risk of contextual DQ issues developing.

To understand the difference between an intrinsic characteristic and a contextual characteristic, consider that mass is a physical intrinsic property of any physical object, weight is a contextual property that varies depending on the strength of the gravitational

field in which the respective object is placed. When DQ is perceived as fitness for use, the DQ is defined contextually by the user's requirements. CDQ issues include logical inconsistency of information and improbable distribution of data element values.

It is generally inadvisable to make assumptions about a dataset derived from another [11]. Past study of electronic health record (EHR) DQ has revealed highly variable results. Hogan and Wagner [15] in their 1997 literature review found that the accuracy of data ranged between 44% and 100%, and completeness between 1.1% and 100%, depending on the clinical concepts being studied. In a medical setting, DQ issues may lead to decreased care quality, introduce privacy and other civil liberty concerns, create liability risks, undermine the reliability and benefits of information technology (IT) investments, deter adoption of health IT, and cost lives [16]. Systematic methods of assessing the quality of an EHR-derived dataset for subsequent research tasks are needed.

While traditional DQ efforts primarily examine data elements intrinsically, data collected for translational research studies integrated with data collected as part of patient care is at risk of CDQ issues. Integrated data has the potential to be more informative with respect to DQ than the sum of its isolated elemental constituents. One data element can be cross-checked against another for logical consistency within the context of co-existing data elements with a data set, or used as raw data to calculate new data elements.

THE IDQ FRAMEWORK

Missing fields and missing values are among the subtypes of the intrinsic data quality (IDQ) threat incompleteness. When rows from two different data sources are concatenated into one table, and one data source contains a column that is not contained in the other data source, the resulting integrated data will have a missing field for each row that originated from the data source that could not supply the field. A missing value occurs when a field (storage for a variable value) exists, but no data value is stored for the variable in an observation. Possible causes of a missing value include:

- Programming error
- Inadvertent data entry omission
- Optional field
- Nonconformance regarding required fields
- Data entry is pending

Both missing fields and missing values result in incompleteness, i.e. missing data. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. All data elements are vulnerable to incompleteness. Missing data can usually be programmatically detected by querying for "Null" values.

Other intrinsic DQ threats involve data which exist but is inaccurate. An inaccuracy occurs when a field exists, and a value is stored for the variable in an observation, but the statement is in fact erroneous. Inaccuracy threats include:

- Date of actual occurrence is in the future
- Numeric value out-of-specified-range
- Categorical value not part of a specified set of allowable values
- Formatted text value does not conform to specified format
- Generally erroneous data

Data-type-based threats can usually be programmatically detected by queries based on established business rules. However, IDQ assessment techniques do not effectively detect the last threat, erroneous unformatted text data. Possible causes of inaccurate data include:

- Programming errors
- Data entry error
- Differing units within the same lab test
- Preliminary data entry to be subsequently updated
- Data dictionary is lacking or not available
- Data dictionary is not adhered to
- Guidelines are not adhered to

- Unclear data definitions
- Unclear data collection guidelines

THE CDQ FRAMEWORK

Where intrinsic DQ assessment techniques fail, CDQ assessment techniques may succeed. A contextual perspective may expose a data value which intrinsically appears true, but is not. Any combination of the previously listed causes may result in threats to CDQ threats. Data profiling of integrated data uniquely enables identification of *contextual* DQ threats including logical inconsistency, low cardinality, and anomalous distribution.

A logical inconsistency is logical incompatibility or contradiction between two or more statements. Each value of two associated attributes may appear accurate when examined intrinsically, but juxtaposed may expose previously undiscovered problems. Despite the absence of any DQ issues in the individual data elements, their integration can create significant errors resulting from subtle and minor discrepancies in business rules between multiple sources. For example, while each data element has a specific set of values prior to integration, a new specified set of values is created for the composite data element. Thus, an opportunity exists to query based on those business rules to test for logical inconsistency. Chiang and Miller call these conditional functional dependencies (CFDs) [17].

For a category to be useful, it should contain a minimum number of records. The more categories there are, the more information there is available for review and summarization [18]. However, too many categories, while adding complexity, may not add useful detail. Although the number of categories that is too many may be subjective, each categorical data element that does not exhibit low-cardinality potentially contains too many categories. Low-cardinality refers to columns with few unique values. Low-cardinality column values are typically status flags, Boolean values, or major classifications such as gender. Thus, a low cardinality violation occurs when a categorical data element contains a significant quantity of unique values.

Low cardinality alone does not guarantee high utility. One or two records may include something quite intriguing, but if they are one or two out of 1,000 records, the information they contain may not be frequent enough in the population to be practically useful [18]. Distribution analysis of categorical data may reveal data that threatens value not with inaccuracy, but with usability challenges. Distribution analysis of numeric data elements can often be compared to an expected distribution. Such comparison might reveal that a data value which intrinsically appears accurate, contextually appears as an outlier or anomaly and worthy of explanation. An investigation may lead to detection of a data error, refinement of business rules, or scientific discovery. A numeric data value may appear as an outlier or as a data statistical distribution anomaly, when compared to the entire available population of values for the same data element. A data anomaly is a

value or set of values failing to conform to an expected (or useful) pattern [19], e.g. a statistical distribution pattern.

Distribution analysis is not typically informative regarding data that are not numeric or categorical. For each numeric or categorical data element, however, a histogram of its population of values can be created to visualize its distribution and visually detect outliers and nonconformance to the expected pattern. The Pareto principle (also known as the 80–20 rule, the law of the vital few, and the principle of factor sparsity) states that, for many events, roughly 80% of the effects come from 20% of the causes [20]. Usability may be better supported when:

- 80% of an element's instances are attributed to 20%+ of the categories
- 80% of an element's categories contribute to 20%+ of the instances

Partitioning each histogram in quintiles to apply the Pareto principle may help to expose perceived defects and outliers. The Pareto Principle, however, is not an established scientific theorem, but rather an operational heuristic. Researching different partitioning methods may be beneficial.

The goal of our work is to apply this simple and easily reproducible framework to examine the value of datasets with intrinsic and contextual methods. The hypothesis of this study is that in addition to traditional/standard methods to measure IDQ, the application of additional CDQ methods to ANK related data will detect more DQ issues.

THE DATA SOURCE

This study focused on the oncology domain and the calculation of natural killer cell counts as the context. If the absolute lymphocyte count (ALYM) of a blood sample is known, and the percentage of those ALYMs which are natural killer (NK) cells are known, the absolute count of NK cells (ANK), an outcome of many oncology clinical trials, can be calculated (ANK [cells/uL blood] = ALYM [cells/uL blood] x % of lymphocytes which are NK cells). NK cells are lymphocytes which play an important role in the innate immune response to infection and cancer and are studied extensively as potential therapeutic agents. The data was obtained from the Masonic Cancer Center (MCC) which is one of 41 certified comprehensive cancer centers nationwide, and one of two in Minnesota, as recognized by the National Cancer Institute (NCI) (National Cancer Institute). MCC's Oncology Medical Informatics and Services group (OMIS) develops and maintains MCC's repository of oncology clinical trial integrated data called the Oncology Business Layer (OBL). The OBL has accumulated health data for more than 2,000 clinical research protocols and 10,000 subjects treated for cancer as early as 1968. Data from clinical, protocol management, flow cytometry, molecular cell therapy, bio-repository, and bone marrow transplant information systems are copied nightly to a landing area where general extract/transform/load processes are run to populate the OBL.

The OBL is where the raw data necessary to calculate ANK is first integrated. The data collection process for each subject begins when a cancer patient at a hospital

(typically at the University of Minnesota Medical Center Fairview, but sometimes another institution) is enrolled as a subject in one or more MCC clinical research trials. A hospital typically collects blood samples from each patient (typically daily but sometimes more frequently) for complete blood count (CBC) testing as part of patient care and without regard to any research protocol. Included in CBC results are those for ALYM and those for white blood cell (WBC) count. For each patient who is also a subject, available test results are electronically transmitted to the OBL. For each subject, at specific time points determined by a research protocol calendar, a separate processed blood sample associated with a Patient Identifier and a Collection Date is sent to MCC's research laboratory for additional testing. The processed sample cannot be used to determine ALYM or WBC levels. The research laboratory's test results are also electronically transmitted to the OBL, and include the % of lymphocytes which are NK cells (%NK). To calculate ANK, for each %NK numeric test result, its associated Patient ID and Collection Date pair are used to match to ALYM and WBC results to link each %NK result to all ALYM and WBC results collected for that patient on that date. MCC's total populations of 6442 ANK blood sample collection records were this obtained over a period from 2003 to 2013.

APPLICATION OF THE IDQ FRAMEWORK TO ANK DATA

IDQ assessment methods listed in Table 1 were applied to each variable; %NK, Collection Date, Patient ID, and ALYM to identify IDQ issues. Each variable was also evaluated for completeness. The value of each numeric data element (in this case %NK and ALYM) was examined for whether it fell within an expected range. If it fell outside of the range, both the numeric value and record containing the value was marked as having an IDQ issue. Because %NK is a value representing a percentage, the expected range for %NK must be from 0 to 100. The expected range for ALYM was 0 to 1054 per mL of blood, the largest value recorded. Collection Date was tested for a date which was logical. If the date was in the future, both the date and the record containing it was marked as having an IDQ issue. Patient ID (text) was examined to determine whether it was in the expected format for MCC. The expected format of the patient identifier was either a ten-character hospital medical record number (MRN) or an OMIS-generator identifier containing the characters “KIR.”.

	IDQ Framework		
Data Element Name	Data Type	IQD Test Type	IDQ Data Element Test Rule
%NK	Numeric	Valid range	0-100
Collection Date	Date	Logical date	Not future-dated
Patient ID	Text	Consistent format	10 characters long or contains characters “KIR”
ALYM	Numeric	Valid range	0-1054

Table 1. The IDQ Framework applied to ANK data.

APPLICATION OF THE CDQ FRAMEWORK TO ANK DATA

The CDQ framework was used to develop and apply CDQ methods for each data element (Table 2). Each same row of data was further evaluated and flagged as containing a CDQ issue, if it had a calculated ANK result outside the range of 0 to 1054, an ALYM result greater than its associated WBC result, and/or no ALYM result paired with an existing %NK result. Descriptive statistics for CDQ versus IDQ issues are presented in the Results section.

	CDQ Framework	
Data Element Name(s)	CQD Test Type	CDQ Case Rule
ALYM, WBC	Logical Consistency	WBC >= ALYM from the same blood sample collection
%NK, ALYM	Logical Consistency	At least one ALYM exists when a %NK exists for the same patient on the same collection date
ANK=%NK*ALYM/100	Valid Range	0-1054

Table 2. The CDQ Framework applied to ANK data.

STUDY RESULTS

	Records exposed contextually, but not intrinsically	Records exposed intrinsically, but not contextually	All records that were exposed intrinsically	All records that were exposed contextually	All records
IDQ Results					
Count of %NK	3177	0	1161	4338	6442
Count of %NK Out-of-Range	0	0	0	0	0
Count of Patient ID	3177	0	1161	4338	6442
Count of Patient ID Inconsistent Format	0	0	720	720	720
Count of Collection Date	3177	0	1161	4338	6442
Count of Collection Date Invalid	0	0	0	0	0
Count of ALYM	1	0	444	445	2549
Count of ALYM Out-of-Range	0	0	444	444	444
Count of Records with at least 1	0	0	1161	1161	1161
CDQ Results					
Count of WBC	453	0	0	451	2552
Count of ALYM>WBC	1	0	0	1	1
Count of Missing %NK link to ALYM	3176	0	717	3893	3893
Count of ANK	1	0	444	445	2547
Count of ANK Out-of-Range	0	0	444	444	444
Count of Records with at least 1	3177	0	1161	4338	4338

Table 3. Complete DQ results.

The CDQ Framework revealed 3177 DQ issues above and beyond what could be detectable through intrinsic methods. The majority (3176) of these issues were contextually exposed by cross-referencing each %NK value with its associated ALYM

value and flagging an error if unsuccessful. The other 1 was contextually exposed by cross-referencing each ALYM value with its associated WBC value (Table 3).

No issues were exposed intrinsically, that were not also exposed contextually. After applying the IDQ Framework to the ANK data, no result for %NK was found to be beyond the valid range (0-100), and no Collection Date was found to be a date in the future. However, 720 Patient IDs were found to be malformed, and 444 ALYM values were found to be beyond the valid range, 0-1054. Using the IDQ Framework, a total of 1161 out of 6442 (18%) sample collection records were flagged has having at least one IDQ issue. The same 1161 records were flagged contextually through two distinct CDQ tests. 444 ANK calculations were beyond the valid range, 0-1024, and 717 [%NK] collections lacked an association to an ALYM test result (Table 3).

Using the IDQ Framework, a total of 1161 out of 6442 (18%) sample collection records were flagged has having at least one IDQ issue. Applying the CDQ Framework to the same 6442 records, 4338 (67%) records were flagged as having at least one issue by 3 distinct CDQ tests. 446 ANK calculations were beyond the valid range (0-1024), 3 ALYM results were erroneously greater than their associated WBC result, and 3893 [%NK] results lacked an association to at least one ALYM test result (Table 3).

DISCUSSION

CBCs (including ALYM and WBC) are processed in a hospital from MCC clinical trial subjects as part of routine patient clinical care. An evaluation of this clinical data

would reveal that the data was complete (i.e. containing no missing values) when all results were found for all tests that should have been taken according to the current prevailing standard of care.

Separately, immune monitoring tests (including %NK) are processed in MCC's research laboratory for each of the same MCC clinical trial subjects as determined by a specific research protocol. An evaluation of this research data would reveal that the data was complete when all results were found for all tests that should have been taken as specified by each test's governing protocol.

CDQ assessments identify issues with integrated data that cannot be found intrinsically

The majority (3176) of records flagged contextually that were not flagged intrinsically were those where a %NK value existed, but an associated ALYM value did not exist. Only after integrating the ALYM data with the %NK data from the research lab, can this be evaluated in the context of ANK calculation. This study therefore considers this DQ issue one that is contextual. Possible reasons for the missing ALYM values include incorrectly entered subject identifiers in the system which sources the %NK values preventing a linkage the ALYM value in the clinical source, or a problem with the integrity of transmission of the source system to the target destination, the OBL.

The remaining one issue was contextually exposed by cross-referencing each ALYM value with its associated WBC value and testing the values to see whether the ALYM

was less than or equal to its associated WBC. For both values to be accurate, this must be the case, because lymphocytes are a subset of white blood cells. However, this ALYM of 0.4 was associated with a WBC of 0.2. 0.4 and 0.2 are valid values intrinsically for ALYM and WBC respectively, yet one or both values is clearly erroneous when juxtaposed, making this case a classic example of the need for contextual testing. Data entry error is the likely cause of this CDQ issue.

Issues exposed intrinsically are also exposed contextually

Pre-study expectations were that intrinsic methods would identify some DQ issues, contextual methods would identify more, but the two sets of results would be overlapping. In other words, it was expected that intrinsic methods would detect issues that would not be detected contextually. Surprisingly, no issue was found intrinsically that was not also found contextually.

444 ALYM values were intrinsically found to be beyond the valid range, 0-1054. While most CBC data was sources from Fairview, a subset of CBC data was sourced from one other source. All 444 out-of-range ALYM values were found to be from this source. Differing rules between the two sources regarding the units used for ALYM accounted for discrepancy. Fairview entered ALYM values in cells/uL blood, while the ALYM values from the non-Fairview source appeared to be entered in cells/mL blood resulting in numbers 1,000,000x greater than they should have been. For each %NK result, its associated Patient ID sourced from the research laboratory was intrinsically

tested. 720 Patient IDs were found to be in a format other than what was expected at MCC. Data entry error is a possible cause. The impact of a poorly formatted Patient ID is the potential inability to link to CBC data which is from a source other than the research laboratory. Only 3 of these IDs successfully matched one associated with an ALYM result, but the 3 ALYM results were among those the 444 that were intrinsically found to be out-of-range. Thus, 444 and an additional 717 records were flagged intrinsically for a total of 1161.

The same 444 records that were flagged intrinsically were also flagged contextually by testing each ANK value to see whether it was within the expected range, 0-1054. 717 failed the CDQ test for linking each %NK results to one or more ALYM result. Thus, the same 1161 intrinsically flagged records were also flagged contextually.

CONCLUSION

CDQ assessments identify more issues with integrated data

Within the domain of translational oncology research, the importance of assessing the intrinsic quality of electronic health data is well recognized. With increasing exchange and secondary use of data, now more than ever segmented information must be integrated from multiple entities. Therefore, the primary goal of this study was to explore the hypothesis that the addition of contextual methods of assessing the quality of translational oncology research data is more effective than intrinsic methods alone. This study

developed simple frameworks to apply traditional IDQ measures and to create novel CDQ measures for oncology research data elements. Each of the two frameworks was then applied the same test dataset, using calculated ANK counts, to compare the effectiveness of the two approaches to DQ. In contrast to intrinsic methods alone, the addition of contextual methods increases the number of DQ issues detected. Furthermore, the CDQ framework captured 100% of the issues that were identified by the IDQ framework. This example demonstrates that the framework can be applied to translational oncology research data to enhance the quality and ultimately support better research for the development of new treatments for cancer.

Limitations to this study include using one case study, examining of a small number set of data elements, and the utilization of no more than three data sources. The data examined within this study also did not test all the components of the proposed frameworks including examining data element distribution anomalies and categorical data. Finally, results were not analyzed in terms of dimensions such as time, gender, or disease. These limitations were useful in meeting the study goal of developing and implementing a simple and reproducible pair of DQ frameworks. Future work may address these limitations, while building upon the work presented here.

Traditional DQ efforts are intrinsic in nature. This represents a gap in the domain that is worthy of additional exploration. Global analysis, visualizations, and descriptive statistics represent a potentially powerful framework for assessing the quality of

integrated oncology research data, and ultimately improving patient care. DQ issues within oncology data collected for translational research combined with secondary use of data collected during clinical treatment may better be detected with novel methods that take into consideration the context.

CHAPTER 3: PROFILING VS. SME RECALL IN DEFINING RULES FOR TRANSLATIONAL CANCER RESEARCH DATA QUALITY

Giordi Orreggio, MHI[1], Sarah Cooley, MD[2], Stuart Speedie, PhD[1]

[1]Institute of Health Informatics, Minneapolis, MN

[2]Masonic Cancer Center, University of Minnesota, Minneapolis, MN

SYNOPSIS

Translational cancer research relies heavily on data collected longitudinally about subject visits and subsequent clinical outcomes. The context of the single data point is often not yet defined at the time of data entry, which raises the risk of introducing conditional functional dependency (CFD) inconsistencies. Many tools for constraining data using rules to detect CFD inconsistencies exist, but little guidance is available regarding how to determine such rules. We compared the results of CFD rules generated by subject matter experts (SME) recall to CDF rules generated by data profiling. Data profiling identified three more CFD rules than the seven previously defined by subject matter expert (SME) review. Since confirmation of previously identified CFDs motivated reanalysis and rule modification, data profiling appears to be useful for confirming data rules documented by SMEs. Using both SMEs and data profiling results

in a better understanding of rules for determining data quality of translational cancer research data.

SIGNIFICANCE

Past studies of electronic health record (EHR) data quality (DQ) have revealed highly variable results. Hogan and Wagner [15] in their 1997 literature review found that the accuracy of data ranged between 44% and 100%, and completeness between 1.1% and 100%, depending on the clinical concepts being studied. In a medical setting, DQ issues (DQIs) may lead to decreased care quality, introduce privacy and other civil liberty concerns, create liability risks, undermine the reliability and benefits of information technology (IT) investments, deter adoption of health IT, and cost lives [16].

Data rules can be used to test and evaluate data quality. Data rules provide a method to define specific tests (i.e. validations and constraints) associated with data, and identify exceptions to expected conditions. Such rule application or tests may evaluate to a true or false value to set up pass or fail checks to assess DQ. They represent logical expressions that can include multiple conditional expressions, and they can contain simple or complex and nested Boolean conditions [21].

Current DQ improvement efforts tend to focus on accuracy [9] [22] [23] and completeness [24] [13]. Accuracy-related measures such as precision, value within range, and allowed values, and completeness-related measures such as missing values are

typical measures of DQ [24] [13]. We label these intrinsic measures. Because our area of research, translational cancer research, relies heavily on data collected over time about subject visits and subsequent clinical outcomes, intrinsic data quality (IDQ) approaches alone that focus on individual data items independent of all others may fall short of characterizing the quality of the data.

For example, when an instance of a required data element called Gender is "Male" and an instance of a required data element called Cancer Site is "Ovary" for the same patient, while the values are both valid and not missing, they are logically inconsistent with each other, given the rule that males are not known to have ovaries. Chiang and Miller [17], and Bohannon and Fan [25] among others, call these conditional functional dependencies (CFDs). They believe that in contrast to traditional functional dependencies that were developed mainly for schema design, CFDs aim at capturing the consistency of data by incorporating bindings of semantically related values. This not only yields a constraint theory for CFDs but is also a step toward a practical constraint-based approach for improving DQ.

PREVIOUS STUDY

Our previous study of contextual vs. intrinsic evaluation of DQ [26] demonstrated that translational cancer research data element instances appearing to have no intrinsic data quality issues (IDQIs) could have conditional functional dependency inconsistencies

(CFDIs), a contextual data quality issue. In that study, data rules were defined directly using subject matter expert (SME) recall [27]. Our CFDI framework defined a simplified process of pairing each candidate data element with another candidate data element and relying on a SME to determine whether one or more testable CFDs are applicable.

As a validation process, results from these preliminary tests can then be analyzed to determine whether it an actual potential error, or an effect of a misunderstanding of the associated rules. When a detected CFDI is deemed to be a false positive, a new understanding of a rule is gained, associated tests can be modified, and the process can be repeated until all the rules applied appear plausible and valid.

For example, in our previous study, users of blood and bone marrow transplant (BMT) cellular product infusion data generally believed that an acceptable intrinsic rule for the Recipient BMT Identifier (ID) is that it is to be formatted as an 8 characters value, beginning with 4 digits and ending with 4 alphabetical characters. Based on this rule, initial testing of Recipient BMT ID resulted in 108 "malformed" values. However, upon further examination, patterns in these Recipient BMT IDs suggested intentional deviation from a default format, rather than entry errors. After reiteration and refinement for the final analysis a modified IDQ rule and also an additional CFD rule were applied, each more complex than the initial IDQ rule. Similarly, the rule for Donor BMT ID is believed to be a value that is "null" or 8 characters long, beginning with 4 alphabetical characters and ending with 4 digits. Based on that rule, the original IDQ testing for

Donor BMT ID resulted in 70 “malformed” Donor BMT IDs. After initial inspection, however, the original simple initial IDQ rule for Donor BMT ID was replaced by a different IDQ rule and an additional CFD rule.

In our previous study, we defined logical inconsistency as logical incompatibility or contradiction between values of two or more data elements. Our CFDI testing conceptually constrains one or more values of one data element to one or more values of another data element within the same case, testing logical consistency and enumerating CFDIs. We defined a simple process of pairing each data element that is a candidate for examination with every other candidate data element and then, for each pairing, relying on existing subject matter expertise to determine whether one or more testable CFDs for that pair existed. Each resulting CFD was incorporated into a rule for logical inconsistency regarding that data element pair. This resulted in three SME-identified CFDs that paired categorical data as shown in Table 1.

Conditional Functional Dependency Inconsistency (CFDI)	Rule
inconsistent ***Donor Relationship*** -> ***Product Relationship***	"Self" -> "Autologous" "Not entered" -> "Not entered" Any other -> not "Autologous"
inconsistent ***Product Type*** -> ***Donor CMV Status***	A cord blood -> not "Positive" Any other -> any
inconsistent ***Product Type*** -> ***Product Relationship***	"Not entered" -> "Not entered" A cord blood -> "Allogeneic" “NK (Natural Killer) Cells” -> Allogeneic “Cellerant Therapy” -> Allogeneic Any other -> not "Not entered"

Table 4: Our CFDI framework applied in our previous study producing three CFDIs

BACKGROUND

While our previous study suggested that this approach was effective at increasing both the understanding of data element pair relationships and the detection of CFDI, it did not account for the possibility that not all rules were discovered. An alternative to relying on SME recall is to expose relationships between data element pairs from data profiling results [27]. One data element's dependency on another influences the distribution of values within the two data elements. For categorical data elements, this effect can be meaningfully quantified since the number of discrete categories is generally small. It is less applicable to data elements with a large range of values such as dates, identifiers and numeric measures.

Consider an example where the unique values of a data element called Cancer Site {*Ovary*, *Prostate*, and *Lung*} and the unique values of a data element called Gender are {*Female* and *Male*}. The minimum number of possible unique combinations of the two data elements is the number of unique data element values (cardinality) of whichever set is larger, in this case, 3. Each value of the first data element paired with each value of the second data element produces the maximum number of possible unique combinations, in this case, 6. In the absence of a rule linking the two component data elements, we would expect to observe a number of unique combinations closer to the maximum. On the other hand, given the rule that {*Prostate*} applies to males only and {*Ovary*} to females only,

we would expect to observe 4 actual unique combinations. This count of 4 on a scale of 3 to 6 where 3 is 0% and 6 is 100% can be represented as 33%, i.e. the two data elements are more dependent than not.

Although two data elements may have a degree of dependency, an extremely low independence value (as calculated above) does not always suggest the existence of a testable rule. Extremely low independence is also seen when a data element with high cardinality is involved. High cardinality refers to the situation where a data element's values tend to be unique as in numeric data, ordinal data, a date or an ID. For example, observing the pairing of one value a data element to one ID or a minority of IDs within a dataset is more likely than observing the pairing of the one value to all IDs or most IDs within the dataset. Thus, our distribution test of independence is designed to be effective at suggesting the existence of a testable data rule in categorical data only.

Similar cross-tabulation techniques are used by Cramer's V to test the independence of two categorical data elements, and by other chi-square-related statistics. However, there are numerous limitations of these statistics. Some require a minimum sample size, or have a maximum number rows and columns in a table. Others lack comparability between tables of different sizes. Most require some statistical expertise. Out technique is simpler to use and appears to detect the existence of data dependencies that can be used to generate rules to some degree.

We attempted to answer the research question, "Is information about data rules improved when a computational approach involving data profiling is added to SME recall?" Our main experiment, therefore tests the hypothesis that translational cancer research data element instances that appear to have no CFDIs from a subject matter expert's perspective can have CFDIs exposed by such computational approaches.

METHODS

Step 1 – Select the data targeted for analysis

The data were obtained from the Masonic Cancer Center (MCC) which is one of 41 certified comprehensive cancer centers nationwide, and one of two in Minnesota, as recognized by the National Cancer Institute (NCI) [28]. MCC's Oncology Medical Informatics and Services group (OMIS) develops and maintains MCC's repository of oncology clinical trial integrated data called the Oncology Business Layer (OBL). The OBL has accumulated health data for more than 2,000 clinical research protocols and 10,000 subjects treated for cancer since 1968. The OBL is updated daily with data from clinical, protocol management, flow cytometry, molecular cell therapy, bio-repository, and blood and bone marrow transplant (BMT) information systems.

As defined in Table 5, we focused on the categorical data elements reported for all cellular product infusions. These data contain a representative sample of the kind of information that is often most important to the translational cancer researcher, patient

clinical outcomes and factors which are suspected to contribute to those outcomes. The combination of a ***Recipient BMT identifier (ID)*** and ***Infusion Number*** uniquely identifies each cellular product infusion.

Data Element Name	Data Element Definition
Diagnosis Type	Name of the disease (the first local diagnosis)
Donor CMV Status	Serostatus result of donor collected prior to transplant of test for cytomegalovirus-specific IgG showing prior exposure to CMV
Product Type	Category of cellular product being infused
Product Relationship	Category of cellular product in terms of genetic sameness to recipient
Recipient CMV Status	Serostatus result of recipient collected prior to transplant of test for cytomegalovirus-specific IgG showing prior exposure to CMV
Donor Relationship	Relationship of donor to the recipient

Table 5: The categorical data elements associated with all cellular product infusions

Step 2 – Count the number of unique values found for each data element

We counted the unique values existing for each of the categorical data elements, so that we could percentages of independence.

Step 3 – For each pair of data elements, calculate the percentage of independence

We paired each of the six selected cellular product infusion data elements with each other and then:

- Determined the minimum possible number of unique combined values resulting from each pairing (the largest count of the two component data elements)
- Counted the actual unique combined values resulting from each pairing
- Determined the maximum possible number of unique combined values resulting from each pairing by multiplying the two component data elements
- For each pairing, applied an algorithm to calculate the two component data elements' percentage of independence defined as the standardized measure of the actual unique combined values on a scale ranging from 0%, representing the minimum possible number of unique combined values to 100% representing the maximum possible number of unique combined values.

Step 4 – Compare percentage of independence with functional dependencies previously defined by SMEs

We predicted that each data element pair both involving only categorical data elements and appearing to be more dependent than not (scoring less than 50%

independent) would have a CFD that could be associated with an explanatory rule. For such cases, we brought the relationships to the attention of an SME for additional review to either define a rule if it did not previously exist, or if it did, confirm or modify the rule.

RESULTS

As reported in

Data Element 1 Data Element 2 Data Element Pair

Field 1 Name	Field 1 Distinct Value Count	Field 2 Name	Field 2 Distinct Value Count	Pair Minimum Possible Distinct Value Count	Pair Actual Distinct Value Count	Pair Largest Possible Distinct Value Count	% of Independence
Donor Relationship	14	Product Type	18	18	72	252	23
Donor Relationship	14	Product Relationship	6	14	33	84	27
Product Relationship	6	Product Type	18	18	49	108	34
Diagnosis Type	16	Donor Relationship	14	16	101	224	41
Diagnosis Type	16	Product Type	18	18	145	288	47
Donor CMV Status	5	Product Type	18	18	52	90	47
Diagnosis Type	16	Product Relationship	6	16	55	96	49

The element pairs above are dependent, while the pairs below are independent.							
Donor Relationship	14	Recipient CMV Status	5	14	42	70	50
Donor CMV Status	5	Donor Relationship	14	14	44	70	54
Product Type	18	Recipient CMV Status	5	18	63	90	63
Diagnosis Type	16	Recipient CMV Status	5	16	57	80	64
Diagnosis Type	16	Donor CMV Status	5	16	58	80	66
Donor CMV Status	5	Product Relationship	6	6	22	30	67
Product Relationship	6	Recipient CMV Status	5	6	22	30	67
Donor CMV Status	5	Recipient CMV Status	5	5	20	25	75

Table 6, the categorical data element pairs scored within the range from 23% to 75% for independence:

- 4 completely new CFDs were discovered (shown in

- Data Element 1
- Data Element 2
- Data Element Pair

Field 1 Name	Field 1 Distinct Value Count	Field 2 Name	Field 2 Distinct Value Count	Pair Minimum Possible Distinct Value Count	Pair Actual Distinct Value Count	Pair Largest Possible Distinct Value Count	% of Independence
Donor Relationship		Product Type		18		252	23
Donor Relationship		Product Relationship		14		84	27
Product Relationship		Product Type		18		108	34
Diagnosis Type		Donor Relationship		16		224	41
Diagnosis Type		Product Type		18		288	47
Donor CMV Status		Product Type		18		90	47

Diagnosis Type		**Product Relationship**		**16**		**96**	**49**
The element pairs above are dependent, while the pairs below are independent.							
Donor Relationship		Recipient CMV Status		14			50
Donor CMV Status		Donor Relationship		14			54
Product Type		Recipient CMV Status		18			63
Diagnosis Type		Recipient CMV Status		16			64
Diagnosis Type		Donor CMV Status		16			66
Donor CMV Status		Product Relationship		6			67
Product Relationship		Recipient CMV Status		6			67
Donor CMV Status		Recipient CMV Status		5			75

Table 6 in bold with gray background)

- 3 others were confirmed modified (shown in

Data Element 1 | Data Element 2 | Data Element Pair

Field 1 Name	Field 1 Distinct Value Count	Field 2 Name	Field 2 Distinct Value Count	Pair Minimum Possible Distinct Value Count	Pair Actual Distinct Value Count	Pair Largest Possible Distinct Value Count	% of Independence
Donor Relationship		Product Type		18		252	23
Donor Relationship		Product Relationship		14		84	27
Product Relationship		Product Type		18		108	34
Diagnosis Type		Donor Relationship		16		224	41
Diagnosis Type		Product Type		18		288	47
Donor CMV		Product Type		18		90	47

Status							
• Diagnosis Type	**•**	**• Product Relationship**	**•**	**• 16**	**•**	**• 96**	**• 49**
•							
• *The element pairs above are dependent, while the pairs below are independent.*							
•							
• Donor Relationship	•	• Recipient CMV Status	•	• 14	•	•	• 50
• Donor CMV Status	•	• Donor Relationship	•	• 14	•	•	• 54
• Product Type	•	• Recipient CMV Status	•	• 18	•	•	• 63
• Diagnosis Type	•	• Recipient CMV Status	•	• 16	•	•	• 64
• Diagnosis Type	•	• Donor CMV Status	•	• 16	•	•	• 66
• Donor CMV Status	•	• Product Relationship	•	• 6	•	•	• 67
• Product Relationship	•	• Recipient CMV Status	•	• 6	•	•	• 67
• Donor CMV Status	•	• Recipient CMV Status	•	• 5	•	•	• 75

- Table 6 in bold with white background)

Data Element 1 Data Element 2 Data Element Pair

Field 1 Name	Field 1 Distinct Value Count	Field 2 Name	Field 2 Distinct Value Count	Pair Minimum Possible Distinct Value Count	Pair Actual Distinct Value Count	Pair Largest Possible Distinct Value Count	% of Independence
Donor Relationship	**14**	**Product Type**	**18**	**18**	**72**	**252**	**23**
Donor Relationship	**14**	**Product Relationship**	**6**	**14**	**33**	**84**	**27**
Product Relationship	**6**	**Product Type**	**18**	**18**	**49**	**108**	**34**
Diagnosis Type	**16**	**Donor Relationship**	**14**	**16**	**101**	**224**	**41**
Diagnosis Type	**16**	**Product Type**	**18**	**18**	**145**	**288**	**47**
Donor CMV Status	**5**	**Product Type**	**18**	**18**	**52**	**90**	**47**
Diagnosis Type	**16**	**Product Relationship**	**6**	**16**	**55**	**96**	**49**
The element pairs above are dependent, while the pairs below are independent.							
Donor Relationship	14	Recipient CMV Status	5	14	42	70	50
Donor CMV	5	Donor	14	14	44	70	54

Status		Relationship					
Product Type	18	Recipient CMV Status	5	18	63	90	63
Diagnosis Type	16	Recipient CMV Status	5	16	57	80	64
Diagnosis Type	16	Donor CMV Status	5	16	58	80	66
Donor CMV Status	5	Product Relationship	6	6	22	30	67
Product Relationship	6	Recipient CMV Status	5	6	22	30	67
Donor CMV Status	5	Recipient CMV Status	5	5	20	25	75

Table 6: Percentage of independence between pairs of data elements, previously defined CFDs bolded and new CFDs grayed and bolded

Results show that between each of the 2 data elements in the 3 previously defined CFDs there is low independence (27-47%). Regarding the 4 new CFDs, ***Diagnosis Type*** specifically appeared to be associated with ***Donor Relationship, Product Relationship***, and ***Product Type***. ***Product Type*** appeared to be associated with ***Donor Relationship***. Regarding the two newly modified CFDs, we discovered ***Donor Relationship*** -> ***Product Relationship,*** and ***Product Type*** -> ***Product Relationship our previous research (Table 1)***. After additional SME review considering relationships highlighted by low independence scores, we discovered four new and two modified CFDs as shown in **Table 7**. Note that although ***Donor CMV Status -> Product Type*** is identified as being a dependent relationship in Table 3, it does not appear in Table 4. Only the 6 of 7 relationships in Table 3 where something new was found because of the methods in the paper are shown in Table 4. We already knew about the ***Donor CMV Status -> Product Type*** and we did not find out any additional information regarding this relationship.

New or Modified Conditional Functional Dependency Inconsistency (CFDI)	Rule
Inconsistent ***Diagnosis Type*** -> ***Donor Relationship***	Both data elements are required. If ***Diagnosis Type*** is “ALL CONV TO AML,” or “CLL,” or “CMML,” or “FANCONI’S ANEMIA” then ***Donor Relationship*** is not “Self”. Certain diagnoses can be treated with either autologous (self) or allogeneic (non-self) donor hematopoietic stem cell transplants (HSCT). Other diagnoses, listed here, are not routinely treated with autologous donor HSCT at UMN. Therefore, self (autologous) donors would be unexpected for these diagnoses.
Inconsistent ***Diagnosis Type*** -> ***Product Relationship***	Both data elements are required. If ***Diagnosis Type*** is “ALL CONV TO AML,” or “CLL,” or “CMML,” or “FANCONI”S ANEMIA” then ***Product Relationship*** is not “Autologous”. See above.
Inconsistent ***Diagnosis Type*** -> ***Product Type***	Both data elements are required. If ***Diagnosis Type*** is “ALL CONV TO AML,” or “CLL,” or “CMML,” or “FANCONI’S ANEMIA” then ***Donor Relationship*** is not “Autologous Backup” or a tandem cellular product infusion. The terms "Autologous Backup" or "tandem" refer to autologous (self) products that are collected with the intent to be used as a second or emergency autologous HSCT. For example, back-to-back "tandem" autologous transplants are used to treat multiple myeloma. They would not be used for the diagnoses listed here.
Inconsistent ***Donor Relationship*** -> ***Product Relationship***	Both data elements are required. If ***Donor Relationship*** is "Self" then ***Product Relationship*** is "Autologous" and vice versa. If ***Donor Relationship*** is “Identical Twin” then ***Product Relationship*** is “Syngeneic” and vice versa. Any other ***Donor Relationship*** -> ***Product Relationship*** combination is acceptable. Autologous product comes from the patient (self-donor), and similarly identical or "syngeneic" twins can provide allogeneic products which are characterized by completely identical genetics (not just identical HLA type).
Inconsistent ***Product Type*** -> ***Donor Relationship***	Both data elements are required. If ***Product Type*** is “Autologous Backup” or “Tandem 1st Infusion” or “Tandem 3rd Infusion” then ***Donor Relationship*** is “Self”. If ***Product Type*** is “Cellerant Therapy” then ***Donor Relationship*** is “Unrelated”. The same rationale is used there (Autologous Back up and Tandem products are from self-donors), whereas Cellerant products refer to an allogeneic product tested in a clinical trial.

Inconsistent ***Product Type*** -> ***Product Relationship***	Both data elements are required. If ***Product Type*** is "Autologous Backup" then ***Product Relationship*** is Autologous." If ***Product Type*** is "BMT Infusion" then ***Product Relationship*** is "Allogeneic". If ***Product Type*** is any other except for "Other*" or "Tandem*" then ***Product Relationship*** is not "Allogeneic".

Table 7: Highlighted are modifications to rules based on reanalysis motivated by results of independence testing

DISCUSSION

We set out to find what improvement, if any, was possible for rules when one adds a computational approach involving data profile results to SME recall. The computational approach we developed first measures the percentage of independence one data element has from another, and based on the measure value can suggest a data rule. The maximum score of 100% occurs when all the possible unique pair values are represented in the data set, representing non-attenuated variability, suggesting minimal relatedness and no associated data rule. If half of the possible unique pair values are represented, the score is 50%. The minimum score of 0% occurs when only the minimum number of the possible unique pair values was represented in the data set.

Because we suspected that the categorical data element pairs that seemed more related than not (less than 50% independent), fell into this as the result of an underlying data rule, these data element pairs were reanalyzed by SMEs. The reanalysis led to the discovery of 4 completely new CFDs above and beyond the seven identified by initial SME review, and the modification of two others.

Our main experiment therefore provides evidence that translational cancer research data element instances that appear to have no CFDIs from a subject matter expert's perspective can have CFDIs exposed by such computational approaches. Our independence scores did not seem to occur by chance. When independence was moderately low (23% to 49%), a data element pair consisting of only categorical data elements could be associated to a CFD and a rule could be defined for that CFD. This DQ profiling appears applicable only to categorical data.

Perhaps the combination of a computational method and SME is more powerful than anyone one of the methods alone, because of inherent limitations with each approach. An SME is limited by the observer's experience. Our computational method has no such limitation, but instead can expose patterns without the constraint of assumptions. Our computational approach, however, stops short at just exposing the patterns, rather than explaining why the patterns exist. The implications our results have for DQ is that a combination of both SME and a computational method is a beneficial approach, as each method compensates for the limitations of the other.

The data set was purposefully constrained to allow testing while avoiding confounding the results. This could affect applicability to broader use cases. Specifically, a limitation to our study is that we restricted CFD testing to only simple two-variable relationships. It is not apparent that analyzing the relationship between combinations of greater than two data elements would provide new and non-redundant

information, because each relationship of such a type can often be represented by multiple two-variable relationships, if the distinction between transitive relationships and direct relationships is not important. Consider an example where the unique values of a data element called Cancer Site {*Ovary*, *Prostate*, and *Lung*}, the unique values of a data element called Gender are {*Female* and *Male*}, and the unique values of a third data element called Has Cancer are {*Yes* and *No*}. A three-way rule is that when Cancer Site is {*Ovary*}, Has Cancer is {*Yes*} AND Gender is {*Female*}. This is the same as having the first of two rules stating that when Cancer Site is {*Ovary*}, Has Cancer is {*Yes*} AND the second of two rules stating that when Cancer Site is {*Ovary*} Gender is {*Female*}.

Another limitation is that the data set that we used, cellular product infusions, is not representative of all data. A third limitation is that there may be other types of DQ issues that are not addressed by CDF (e.g. data timeliness). Future work may address these limitations, while building upon the work presented here.

CONCLUSION

We described a CDQ framework combining SME recall and a data profiling approach representing novel contextual measures for oncology research data elements. Our computational approach involving data profile results added to SME recall enabled identification of data element pairs with moderately low independence (23% to 49%), and suggested relatedness if both data elements of a data element pair were categorical.

This in turn motivated SME reanalysis, and successfully led to an improvement in information about rules, proving our hypothesis.

Note that data profiling alone does not lead to data rule definition. The data profiling merely suggests relationships for an SME to subsequently explain in the form of a rule. Yet, for categorical data, our data profiling method combined with SME analysis was more effective than SME recall alone. The combined approach exposed all rules involving pairs of categorical data elements that were defined initially though SME recall alone, in addition to rules that were not found initially.

For non-categorical data, however, SME recall seemed superior. The high cardinality of non-categorical data confounds our measure of independence, causing the approach to be ineffective at suggesting underlying rules.

Using both SMEs and a data profiling approach results in the best understanding of rules in translational cancer research data. Better data rules enable better definition of specific tests (i.e. validations and constraints) associated with data, and better identification of exceptions to expected conditions, which in turn enables better assessment of data quality.

ACKNOWLEDGMENTS

This work was supported in part by NIH P30 CA77598 utilizing the following Masonic Cancer Center, University of Minnesota shared resource: Oncology Medical Informatics Services (OMIS).

CHAPTER 4: AN EFFECT OF INTEGRATION ON TRANSLATIONAL CANCER RESEARCH DATA QUALITY

Giordi Orreggio, MHI[1], Stuart Speedie, PhD[1], Sarah Cooley, MD[2]

[1]Institute of Health Informatics, Minneapolis, MN

[2]Masonic Cancer Center, University of Minnesota, Minneapolis, MN

SYNOPSIS

Traditional data quality (DQ) efforts focus on intrinsic measures, such as accuracy and completeness. Translational cancer research relies heavily on integrated data collected longitudinally from clinical care, research subject visits, and subsequent clinical outcome data for required reporting. The integration of such data potentially increases the ability to detect DQ issues and thereby remediating them. Methods were developed to assess the DQ of 400,897 blood collection record. This paper highlights this specific example to present a novel method of exposing DQ issues. Compared to the proportion of DQ issues detectable (0.05% and 14.82%) in datasets representing the same body of blood collections but reported through two different mechanisms (HL7 and TIDE), the proportion detectable after integrating the two datasets (34.36%) was much greater.

SIGNIFICANCE

Current DQ improvement efforts tend to focus on accuracy [9] [22] [23] and completeness [24][13]. Accuracy-related measures such as precision, value within range, and allowed values, and completeness-related measures such as missing values are typical measures of DQ [24][13]. We label these intrinsic measures. Because our area of research, translational cancer research, relies heavily on data collected over time about subject visits and subsequent clinical outcomes, intrinsic data quality (IDQ) approaches alone that focus on individual data items independent of all others may fall short of characterizing the quality of the data.

Past studies of electronic health record (EHR) data quality (DQ) have revealed highly variable results. Hogan and Wagner [15] in their 1997 literature review found that the accuracy of data ranged between 44% and 100%, and completeness between 1.1% and 100%, depending on the clinical concepts being studied. In a medical setting, DQ issues (DQIs) may lead to decreased care quality, introduce privacy and other civil liberty concerns, create liability risks, undermine the reliability and benefits of information technology (IT) investments, deter adoption of health IT, and cost lives [16].

Data rules can be used to test and evaluate data quality. Data rules provide a method to define specific tests (i.e. validations and constraints) associated with data, and identify exceptions to expected conditions. Such rule application or tests may evaluate to a true or false value to set up pass or fail checks to assess DQ. They represent logical

expressions that can include multiple conditional expressions, and they can contain simple or complex and nested Boolean conditions [21].

For example, when an instance of a required data element called Gender is "Male" and an instance of a required data element called Cancer Site is "Ovary" for the same patient, while the values are both valid and not missing, they are logically inconsistent with each other, given the rule that males are not known to have ovaries. Chiang and Miller [17], and Bohannon and Fan [25] among others, call these conditional functional dependencies (CFDs). They believe that in contrast to traditional functional dependencies (FDs) that were developed mainly for schema design, CFDs aim at capturing the consistency of data by incorporating bindings of semantically related values. This not only yields a constraint theory for CFDs but is also a step toward a practical constraint-based approach for improving DQ.

PREVIOUS STUDIES

One of our previous studies compared contextual vs. intrinsic evaluation of DQ [26] and demonstrated that translational cancer research data element instances appearing to have no intrinsic data quality issues (IDQIs) could have conditional functional dependency inconsistencies (CFDIs), a contextual data quality issue. To find CFDIs, we tested the data for conformance to data rules. A data rule can be defined directly using subject matter expert (SME) recall [27]. An alternative to relying on SME recall is to

derive rules from data profiling results [27]. Rules regarding one data element's dependency on another influence the distribution of values within the two data elements. This effect can be measured.

A limitation of our two previous studies is that the data set that we used, cellular product infusions, was not representative of all data. In this third study, we examine a different data domain, clinical lab tests, to determine more information regarding the applicability of our frameworks to other types of translational cancer research data.

Information domain is not the only variable that may affect DQ. The mechanism of transmission from source to recipient may also impact DQ. With our IDQ, SME-based CFD, and data profiling-based CFD frameworks built, we now have measures to quantify this effect. In our third study, we examine 2 sets of data each from the same source but differing by the mechanism of transmission. We also examine the combined integrated data set. We attempt to answer the research question, "What are the effects of mechanism of transmission and of integration on DQ?" Our main experiment, therefore sought to explore the hypothesis that translational cancer research data element instances transmitted differently will have different DQIs as will integrated data element instances.

CASE STUDY

Clinical research data collection for each subject begins when a cancer patient at a hospital (typically at the University of Minnesota Medical Center Fairview, but

sometimes another institution) is enrolled as a subject in one or more clinical research trials. A hospital typically collects blood samples from each patient (typically daily but sometimes more frequently) for testing as part of patient care and without regard to any research protocol. The Cancer Center's request for an on-going HL7 feed of clinical lab result data from the University of Minnesota Medical Center Fairview for any Fairview patient who was also a Cancer Center research subject began with blood samples collected on 9/21/2006. This clinical lab data from Fairview and at least one other organization is what we refer to as HL7. HL7, most often white blood cell (WBC) and absolute lymphocyte (ALYM) results would be combined with Cancer Center research lab data to answer specific scientific questions asked within the context of Cancer Center research studies.

On multiple occasions, however, during a review of a researcher's combined lab data, the researcher would discover that some *clinical* lab data appeared to be missing. The Cancer Center had a theory that there was confusion regarding when a patient became a research subject, and that this led to a timing issue, which in turn lead to data failing to get transmitted from Fairview. However, efforts to remediate HL7 were unsuccessful.

Many layers of complexity exist within the Fairview clinical lab data. We define a test instance as one test name per patient per physical blood collection date time. We define an observation as one data entry date time when an evaluation of one test instance was conducted. In many cases, a single test instance can have multiple observations

leading to multiple results for the same test instance. In three specific cases, we found that the same observation of the blood test instance had two different results.

In 2010, the Fairview clinical lab data became available via an additional system, which was then called TIDE. TIDE would have clinical results going forward from that time. TIDE also had historical data on blood collection beginning 7/25/2002, data preceding the HL7. Although the source of data for both TIDE and HL7 are the same, the expectation was that TIDE would not have the same unexplained limitations as HL7.

The next step was to validate this assumption. In two previous health informatics studies, we explored an intrinsic vs. contextual data profiling, and SME versus computational contextual data profiling. In this third health informatics study, we use these data profiling approaches to explore using the approaches to expose information about data quality after integrating data. For the Cancer Center, the goal is, for time period that both TIDE and HL7 were available, to support the provision of the most complete and accurate clinical lab data whether it comes from HL7 exclusively, TIDE exclusively, or some combination of both.

METHODS

To setup for our comparison between TIDE and HL7, we extracted the earliest WBC and ALYM results found in both TIDE and HL7 (09/21/2006 5:23 PM), the latest WBC

and ALYM results found in both TIDE and HL7 (06/18/2014 10:59 PM), and all other WBC and ALYM results between the two date/times, along with each associated Patient ID and Collection Date Time. The 4 data elements are described in Table 5.

Data Element Name	Data Element Definition
ALYM	Short for absolute lymphocyte count, this is a measure of the number of lymphocytes (a type of white blood cell) in blood. It is used to evaluate and manage disorders of the blood or the immune system. It is also used to evaluate and manage certain types of cancer and tumors **Invalid source specified..**
Collection Date	Date sample for clinical testing was physically collected. In general, if multiple samples are collected during the same date, only the first sample is used for the research study, as it is the sample that is most likely least influenced by other factors.
Patient ID	Each Subject's primary identifier used to link Subjects across data sourced from disparate systems. It is typically but not always the Fairview medical record number (MRN)
WBC	Short for white blood cell count (leukocyte count), this is usually measured as part of the complete blood count (CBC). White blood cells are the infection-fighting cells in the blood and are distinct from the red (oxygen-carrying) blood cells known as erythrocytes. There are different types of white blood cells, including neutrophils (polymorphonuclear leukocytes; PMNs), band cells (slightly immature neutrophils), T-type lymphocytes (T cells), B-type lymphocytes (B cells), monocytes, eosinophils, and basophils. All the types of white blood cells are reflected in the white blood cell count. The normal range for the white blood cell count varies between laboratories but is usually between 4.3 and 10.8 cells per nanoliter (nL) of blood. A low white blood cell count is called leukopenia. A high white blood cell count is termed **Invalid source specified.**. WBC can be as high as 1,052 **Invalid source specified.**.

Table 8: Data dictionary of our data extract characterizing ALYM-related data elements

We took several steps to aim for having no more than one result per test instance (with one exception noted below in Item 4):

1. We excluded each result that was not the latest observed result for the same instance of a blood test.
2. In the one case where the same ALYM observation of the same blood test instance in TIDE had two different results, neither matching any observation in HL7, we excluded the one which was associated with the inconsistent RESULT_COMPONENT_NAME (one of the observation record attributes) value: {Lymphocytes #}. All other ALYM results in TIDE were associated with a RESULT_COMPONENT_NAME of {Absolute Lymphocytes}.
3. In the one case where the same WBC observation of the same blood test instance in HL7 had two different results, neither consistent with the single observation in TIDE, we excluded the one which was associated with the non-numeric Result value: {Results questioned - new specimen has been requested}.
4. In the one case where the same WBC observation of the same blood test instance in HL7 had two different results, one consistent with the single observation in TIDE and the other inconsistent, we excluded neither, as doing so would be utilizing knowledge that could only be gained by integrating the two datasets, and therefore should be included in this health informatics study results.

To implement the comparison, we profiled three datasets: TIDE, HL7, and the combined TIDEHL7. Per our approach described in our previous studies, after research and consultation with SMEs, we identified intrinsic rules and contextual rules. We

associated WBC and ALYM with intrinsic rules are shown in Table 9.

Intrinsic Data Quality Issue (IDQI)	Rule
Out-of-range or missing ***ALYM***	0-1054
Out-of-range or missing ***WBC***	0-1054

Table 9: Our IDQI framework applied producing of 2 intrinsic data rules

Contextual tests included cross-checking ALYM with WBC. ALYM is a subset of WBC, and therefore should be equal to or less than WBC. Furthermore, the absence of WBC when ALYM exists constitutes a missing data issue. Because ***ALYM*** results can be too small to report, the data element is optional. As an optional data element, testing for missing data would not provide any additional information regarding IDQ. ALYM is a subset of WBC, and therefore related. Finally, in the integrated data set, we flagged any value for the same patient, collection date, and data element as inconsistent, if the value from the HL7 collection process did not match the value from the TIDE collection process.

RESULTS

Within the time frame that collection results were reported through both mechanisms, 09/21/2006 5:23 PM through 06/18/2014 10:59PM, we found 510,396 TIDE observations from 5830 distinct patients. Within the same time frame, we found 274,574 HL7 observations from 3633 distinct patients.

Post-integration (i.e. examining each data set having the context of the other, rather than examining each in a silo) the discrepancy in observation counts is reflected in three DQ measures of HL7: Missing HL7 WBC, Missing HL7 ALYM, and TIDE Collection Missing in HL7. These measures had the greatest counts (268,205 to 268,377). Three analogous measures of TIDE DQ also had high counts (32,145 to 32,556), although not nearly as large. The smallest counts found post-integration were Unequal TIDE-HL7 WBC, and Unequal TIDE-HL7 ALYM, 116 and 92 respectively.

Other DQ measures which could be examined pre-integration remained unchanged post-integration. The measures ranged from 0 for Malformed Patient ID to 4635 Out-of-range HL7 WBCs. Table 10 shows the DQ measures.

	Pre-integration			Post-integration		
Data Source	1: TIDE Only	2: Both TIDE and HL7	3: HL7 Only	1: TIDE Only	2: Both TIDE and HL7	3: HL7 Only
Malformed Patient ID	0	0	0	0	0	0
Out-of-range TIDE WBC	56	46	n/a	56	46	n/a
Out-of-range TIDE ALYM	65	60	n/a	65	60	n/a
Out-of-range HL7 WBC	n/a	99	4536	n/a	99	4536
Out-of-range HL7 ALYM	n/a	79	3	n/a	79	3
TIDE ALYM>WBC	49	33	0	49	33	0
HL7 ALYM>WBC	n/a	56	0	n/a	56	0
Missing TIDE WBC	632	407	n/a	632	407	32547
Missing TIDE ALYM	0	0	n/a	0	0	32145
Missing HL7 WBC	n/a	66	10	268205	66	10
Missing HL7 ALYM	n/a	172	25859	268318	172	25859
Unequal TIDE-HL7 WBC	n/a	n/a	n/a	n/a	116	n/a
Unequal TIDE-HL7 ALYM	n/a	n/a	n/a	n/a	92	n/a
TIDE Collection Missing in HL7	n/a	n/a	n/a	268377	0	0
HL7 Collection Missing in TIDE	n/a	n/a	n/a	0	0	32556

Table 10: More DQI counts are available (highlighted) post data integration

DISCUSSION

Noteworthy DQ measures which could be calculated pre-integration remained unchanged post-integration

Out-of-range issues could generally be explained, and respect to these types of issues, TIDE DQ was generally better than that of HL7. All 102 Out-of-range TIDE WBCs

were attributable to data entry of a single code, 9999999, instead of an actual value, while 4635 Out-of-range HL7 WBCs were attributable to data entry of 49 different text descriptions, instead of an actual value. 118 out of 125 Out-of-range TIDE ALYMs were attributable to data entry of the code, 9999999, and the other 7 to 4 other high values (1166 to 1683) appearing to be data entry errors for 1 patient, while 82 Out-of-range HL7 ALYMSs were attributable to data entry of 10 different text descriptions, instead of an actual value.

ALYM>WBC issues could also generally be explained. 72 of 82 TIDE ALYM>WBC were attributable to the Out-of-range TIDE ALYM described above. The 10 remaining appeared to be data entry errors of ALYM for 5 patients. 40 of 56 HL7 ALYM>WBC were attributable to the Out-of-range HL7 ALYM described above. The 16 remaining were attributable to data entry of a single code, '(Removed)', instead of an actual value for 8 patients.

Noteworthy DQ measures which could be calculated only post-integration

The measures which could be calculated pre-integration oppose the expectation that TIDE and HL7 reflect the same data. A second assumption was that the TIDE data would be more inclusive because it would not have the same unexplained limitations as HL7. The measures which could be calculated only post-integration confirmed the existence of limitations with the HL7 data. The discrepancy in observation counts is

reflected in measures of HL7 DQ -- Missing HL7 WBC, Missing HL7 ALYM, and TIDE Collection Missing in HL7 – informed by comparison to the TIDE data. In this respect, the measures are contextual, as they could only be found only after data integration.

If all HL7 observations were also in TIDE, we might conclude that the HL7 data was of lesser quality than the TIDE data because the HL7 data was less complete, and appropriate action would be to disregard the HL7 and regard TIDE as the source of truth. However, this was not the case. Of the 274,574 HL7 observations, 32,556 from 2404 patients were not matched in TIDE. A limitation of TIDE is also exposed and this limitation is reflected in analogous measures of TIDE DQ -- Missing TIDE WBC, Missing TIDE ALYM, and HL7 Collection Missing in TIDE – informed by comparison to the HL7 data. In this respect, these measures are similarly contextual, as they could only be found only after data integration.

If all 2404 patients of the HL7-only observations were not found in TIDE, we might explore the possibility of a patient selection error within TIDE. Our investigation found that this was not the case. 2156 of the 2404 HL7 patients were also in TIDE.

The smallest counts found post-integration were Unequal TIDE-HL7 WBC, and Unequal TIDE-HL7 ALYM, 116 and 92 respectively. 102 of 116 Unequal TIDE-HL7 WBCs were attributable to either Out-of-Range WBCs, while the remaining 14 appeared to be true discrepancies for 7 patients. 81 of 92 Unequal TIDE-HL7 ALYMS were

attributable to either Out-of-Range ALYMs, while the remaining 11 appeared to be true discrepancies for 6 patients.

CONCLUSION

The primary goal of this study was to explore the hypothesis that translational cancer research data element instances transmitted differently, or integrated will have different DQIs. The results of our study support this hypothesis. Each data collection event can result in missing data from simply utilizing faulty inclusion or exclusion logic. Furthermore, our dataset transmitted through the HL7 mechanism exhibited one primary DQI – out of range ALYM values. Our data transmitted though the TIDE mechanism also exhibited only one primary DQI, but it was a different issue – non-numeric WBC values. Quality check of the integrated dataset was superior. Not only did it reveal both types of DQIs, it also showed the relative contribution of each, but more importantly revealed a greater DQI type – missing data.

It is possible that differences in the processing to create the TIDE data versus the HL7 data account for the differences in the final data sets. Resolving those differences, however, is beyond the scope of this research, which intends to provide a way only to expose the differences in data sets that are believed to reflect the same information.

Limitations to this study include using one case study, examination of a small number set of data elements, and the utilization of no more than three data sources. Future work

may address these limitations, while building upon the work presented here. But for now, it appears that DQIs within oncology data collected for during clinical treatment may better be detected as the level of data integration is increased.

ACKNOWLEDGMENTS

This work was supported in part by NIH P30 CA77598 utilizing the following Masonic Cancer Center, University of Minnesota shared resource: Oncology Medical Informatics Services (OMIS).

CHAPTER 5: DISCUSSION

STUDY ONE: CONTEXTUAL VS. INTRINSIC EVALUATION

The subject of our first study was the calculation of natural killer cell counts. If the absolute lymphocyte count (ALYM) of a blood sample is known, and the percentage of those ALYMs which are natural killer (NK) cells are known, the absolute count of NK cells (ANK), an outcome of many oncology clinical trials, can be calculated (ANK [cells/uL blood] = ALYM [cells/uL blood] x % of lymphocytes which are NK cells). NK cells are lymphocytes which play an important role in the innate immune response to infection and cancer and are studied extensively as potential therapeutic agents.

CBCs (including ALYM and WBC) are processed in a hospital from MCC clinical trial subjects as part of routine patient clinical care. Separately, immune monitoring tests (including %NK) are processed in MCC's research laboratory for each of the same MCC clinical trial subjects as determined by a specific research protocol. An evaluation of this research data would reveal that the data was complete when all results were found for all tests that should have been taken as specified by each test's governing protocol.

The major finding of this study was that our method of detecting DQ issues contextually exposed issues above and beyond those detected by using our intrinsic method. Most records flagged contextually that were not flagged intrinsically were those where a %NK value existed, but an associated ALYM value did not exist. An ALYM result should exist for each %NK result, because the clinical study protocol calls for each

of these two types of lab tests to be performed on the same blood draw. Only after integrating the ALYM data with the %NK data from the research lab can this be evaluated in the context of ANK calculation. Possible reasons for the missing ALYM values include: incorrectly entered subject identifiers in the system which sources the %NK values preventing a linkage the ALYM value in the clinical source, or a problem with the integrity of transmission of the source system to the target destination.

STUDY TWO: A PROFILING VS. SME RECALL APPROACH TO DEFINING DATA RULES

While Study One utilized subject matter expert (SME) recall to identify data rules, in Study Two we automated the process, computationally evaluating each pairing of each data element with every other data element. Note that such data profiling alone does not lead to data rule definition. The data profiling merely may suggest relationships for an SME to subsequently explain in the form of a rule. Yet, for categorical data, our data profiling method combined with SME analysis was more effective than SME recall alone.

The combination of a computational method and SME may be more powerful than any one of the methods alone, because of inherent limitations with each approach. An SME is limited by the observer's experience. Our computational method has no such limitation, but instead can expose patterns without the constraint of assumptions. Our

computational approach, however, stops short at just exposing the patterns, rather than explaining why the patterns exist. The implications our results have for DQ is that a combination of both SME and a computational method is a beneficial approach, as each method compensates for the limitations of the other.

STUDY THREE: AN EFFECT OF INTEGRATION

A limitation of our two previous studies is that the data set that we used, cellular product infusions, was not representative of all data. In this third study, we examine a different data domain, clinical lab tests, to determine more information regarding the applicability of our frameworks to other types of translational cancer research data.

Information domain is not the only variable that may affect DQ. The mechanism of transmission from source to recipient may also impact DQ. With our IDQ, SME-based CFD, and data profiling-based CFD frameworks built, we now have measures to quantify this effect. In our third study, we examine 2 sets of data each from the same source but differing by the mechanism of transmission. We also examine the combined integrated data set. We attempt to answer the research question: "What are the effects of mechanism of transmission and of integration on DQ?" Our main experiment, therefore sought to explore the hypothesis that translational cancer research data element instances transmitted differently will have different DQIs as will integrated data element instances.

CONCLUSION

STUDY ONE: CONTEXTUAL VS. INTRINSIC EVALUATION

The primary goal of this study was to explore the hypothesis that the addition of contextual methods of assessing the quality of translational oncology research data is more effective than intrinsic methods alone. Our results concluded that in contrast to intrinsic methods alone, the addition of contextual methods increases the number of DQ issues detected. Furthermore, the CDQ framework captured 100% of the issues that were identified by the IDQ framework. This example demonstrates that the framework can be applied to translational oncology research data to enhance the quality and ultimately support better research for the development of new treatments for cancer.

There are limitations to this study that could have influenced our conclusion. These include using one case study, examining of a small number set of data elements, and the utilization of no more than three data sources. However, grouping this study with the subsequent two studies addresses this limitation and as such our conclusion stands within the overall context of the studies.

STUDY TWO: A PROFILING VS. SME RECALL APPROACH TO DEFINING DATA RULES

Our second study provides evidence that those translational cancer research data element instances that are categorical but appear to have no CFDIs from a subject matter

expert's perspective can have CFDIs exposed by such computational approaches. For non-categorical data, however, SME recall seemed superior.

In conclusion, using both SMEs and a data profiling approach results in the best understanding of rules in translational cancer research data. Better data rules enable better definition of specific tests (i.e. validations and constraints) associated with data, and better identification of exceptions to expected conditions, which in turn enables better assessment of data quality.

STUDY THREE: AN EFFECT OF INTEGRATION

The primary goal of this study was to explore the hypothesis that translational cancer research data element instances transmitted differently, or integrated will have different DQIs. The results of our study support this hypothesis. Some DQ measures could be calculated only post-integration, which belays any assumption that two distinct data sets deemed fit-for-use continues to be fit-for-use after they are combined.

Limitations to our set of three studies overall include involving using only one organization, The University of Minnesota, and using only two cancer research datasets: a cellular product infusion dataset and a blood draw dataset. Each dataset contained no more than half a dozen data elements. Only three data sources were involved. Our testing seemed applicable to one data type, specifically categorical data, more than others. Finally, results were not analyzed in terms of dimensions such as time, gender, or

disease. These limitations were useful in meeting the study goal of developing and implementing a simple and reproducible pair of DQ frameworks.

Future work may include additional organizations to expose effects attributable to organizational characteristics. Medical data other than cellular product infusion data and blood draw data can also be explored to potentially uncover new findings or reconfirm the findings detailed here, while building upon the work presented here.

Each of these studies clarifies our understanding of different factors in improving cancer research data quality. Data quality efforts are enhanced when contextual methods including cross-checking are added to traditional intrinsic checks such as tests for completeness and accuracy. Using both SMEs and a data profiling approach to review data results in the best understanding of rules in translational cancer research data. Better data rules enable better definition of specific tests (i.e. validations and constraints) associated with data, and better identification of exceptions to expected conditions. When data integration involved, quality checking before and integration provide better results than quality checking before integration alone. Applying these techniques together enables better assessment of data quality.

BIBLIOGRAPHY

[1] S. H. Woolf, "The Meaning of Translational Research and Why It Matters," *JAMA*, vol. 299, no. 2, p. 211–13, January 2008.

[2] National Cancer Institute, "Budget And Appropriations," 23 June 2016. [Online]. Available: http://www.cancer.gov/about-nci/budget. [Accessed 27 August 2016].

[3] NTNU, "What Is Big Data?," 23 June 2016. [Online]. Available: https://www.ntnu.edu/ime/bigdata/what-is. [Accessed 27 August 2016].

[4] CDC, "Number of deaths for leading causes of death," 27 April 2016. [Online]. Available: http://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm. [Accessed 2 October 2016].

[5] C. -. J. Chiang, S. -. L. You, C. -. J. Chen, Y. -. W. Yang, W. -. C. Lo and M. -. S. Lai, "Quality assessment and improvement of nationwide cancer registration system in Taiwan: a review," *Japanese Journal of Clinical Oncology,* vol. 45, no. 3, pp. 291-296, 2015.

[6] D. C. Hsia, "Accuracy of diagnostic coding for Medicare patients under the prospective-payment system," *The New England Journal of Medicine,* vol. 318, no. 6, pp. 352-5, 11 February 1988.

[7] M. Greiver, J. Barnsley, R. H. Glazier, B. J. Harvey and R. Moineddin, “Measuring data reliability for preventive services in electronic medical records,” *BMC Health Services Research,* vol. 12, p. 116, 2012.

[8] S. E. Campbell, M. K. Campbell, J. M. Grimshaw and A. E. Walker, “A systematic review of discharge coding accuracy,” *Journal of Public Health,* vol. 23, no. 3, pp. 205-11, 2001.

[9] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Consumers," *Journal of Management Information Systems,* vol. 12, no. 4, pp. 5-34, 1996.

[10] D. G. Arts, "Defining and improving data quality in medical registries: a literature review, case study, and generic framework," *Journal of the American Medical Informatics Associaction,* vol. 9, no. 6, pp. 600-11, 2002.

[11] N. G. Weiskopf and C. Weng, “Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research,” *J Am Med Inform Assoc,* vol. 20, pp. 144-151, 2013.

[12] H. G. C. F. W. C. Botsis T, "Secondary Use of EHR: Data Quality Issues and Informatics Opportunities.," *AMIA Summits Transl Sci Proc.,* pp. 1-5, 2010 Mar 1.

[13] M. Kahn, "A pragmatic framework for single-site and multisite data quality

assessment in electronic health record-based clinical research," *Medical Care,* vol. 50, no. July, pp. i-ii,S1-S101, 2012.

[14] C. L. Hudson, U. Topaloglu, J. Bian, W. Hogan and T. Kieber-Emmons, "Automated Tools for Clinical Research Data Quality Control using NCI Common Data Elements," *AMIA Jt Summits Transl Sci Proc,* vol. 2014, p. 60–69, 2014.

[15] W. R. Hogan and M. M. Wagner, "Accuracy of Data in Computer-based Patient Records," *Journal of the American Medical Informatics Association,* vol. 4, no. 5, pp. 342-355, 1997.

[16] S. Verhulst, "Background Issues on Data Quality," 2006. [Online]. Available: http://www.markle.org/health/markle-common-framework/connecting-professionals/t5. [Accessed 19 January 2013].

[17] F. Chiang and R. J. Miller, "Discovering Data Quality Rules," *Proceedings of the VLDB Endowment,* vol. 1, no. 1, pp. 1166-1177, 2008.

[18] IBM Corporation, "Tips for Creating Categories," 2011. [Online]. Available: http://pic.dhe.ibm.com/infocenter/spsstafs/v4r0m1/index.jsp?topic=%2Fcom.ibm.spss.tafs.help%2Ftas_cat_create.htm. [Accessed 25 September 2013].

[19] V. Chandola, "Anomoly detection: A Ssrvery," *ACM Computing Surveys,*

vol. 41, no. 3.

[20] N. Bunkley, *Joseph Juran, 103, Pioneer in Quality Control, Dies,* New York City: New York Times, 2008.

[21] IBM, "InfoSphere Information Analyzer," 2009. [Online]. Available: http://pic.dhe.ibm.com /infocenter/iisinfsv /v8r1/index.jsp ?topic=/com.ibm.swg.im.iis.ia.quality.doc /topics/dq_rule_definitions.html. [Accessed 22 July 2014].

[22] D. G. Arts, "Defining and improving data quality in medical registries: a literature review, case study, and generic framework," *Journal of the American Medical Informatics Association,* vol. 9, no. 6, pp. 600-11, 2002.

[23] N. G. Weiskopf and C. Weng, “Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research,” *Journal of the American Medical Informatics Association,* vol. 20, pp. 144-151, 2013.

[24] T. Botsis, G. Hartvigsen, F. Chen and C. Weng, "Secondary Use of EHR: Data Quality Issues and Informatics Opportunities.," *AMIA Summits Transl Sci Proc.,* pp. 1-5, 2010 Mar 1.

[25] P. Bohannon and W. Fan, "Conditional Functional Dependencies for Data Cleaning," *IEEE,* 2007.

[26] S. Speedie, G. Orreggio and S. Cooley, “Intrinsic vs. Contextual Evaluation of Translational Cancer Research Data Quality,” 2014.

[27] Oracle, "Oracle® Warehouse Builder Data Modeling, ETL, and Data Quality Guide," 2011. [Online]. Available: http://docs.oracle.com/cd/E11882_01/owb.112/e10935/data_rules.htm#WBETL19000. [Accessed 22 July 2014].

[28] National Cancer Institute, "NCI-Designated Cancer Centers," [Online]. Available: http://www.cancer.gov/researchandfunding/extramural/cancercenters/find-a-cancer-center. [Accessed 14 August 2013].

[29] American Cancer Society, "Cancer Facts And Figures," 2009.

[30] American Cancer Society, "Cancer Facts & Figures," 2010.

[31] B. A. Kohler, E. Ward, B. J. McCarthy, M. J. Schymura, L. A. G. Ries, C. Eheman, A. Jemal, R. N. Anderson, U. A. Ajani and B. K. Edwards, "Annual Report to the Nation on the Status of Cancer," *J Natl Cancer Inst*, p. 103:1–23, 2011.

[32] M. M. Dr. Sarah Cooley, Interviewee, *Assistant Professor of Medicine; Director, Oncology Medical Informatics; Associate Director, Cancer Experimental Therapeutics Initiative.* [Interview]. 1 March 2012.

[33] Cleveland Clinic, 2012. [Online]. Available: http://www.lerner.ccf.org/qhs/informatics/. [Accessed 20 March 2012].

[34] National Cancer Institute, 2012. [Online]. Available: http://cancercenters.cancer.gov/cancer_centers/index.html. [Accessed 20 March 2012].

[35] Massey Cancer Center, Virginia Commonwealth University, 19 March 2012. [Online]. Available: http://www.massey.vcu.edu/cancer-research-informatics-and-service-core.htm. [Accessed 20 March 2012].

[36] Masonic Cancer Center, University of Minnesota, 19 March 2012. [Online]. Available: http://www.cancer.umn.edu/about/index.html. [Accessed 20 March 2012].

[37] Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services, "Health Information Technology Evaluation Toolkit," [Online].

[38] National Cancer Institute, 14 June 2011. [Online]. Available: http://www.cancer.gov/cancertopics/factsheet/NCI/NCI. [Accessed 19 March 2012].

[39] Office of Cancer Centers of the National Cancer Institute, "Policies and Guidelines Relating to the Cancer Center Support Grant," October 2010.

[Online].

[40] G. Orreggio, "Business Intelligence Dashboards," 12 February 2012. [Online]. Available: https://confluence.cancer.umn.edu/display/OBL/Business+Intelligence+Dashboards. [Accessed 24 March 2012].

[41] University of California, 2012. [Online]. Available: http://accelerate.ucsf.edu/informatics. [Accessed 20 March 2012].

[42] Sarah Frankfurth at Community Clinics Initiative Justin Louie at Blueprint Research and Design, 2005. [Online]. Available: http://healthit.ahrq.gov/portal/server.pt/document/954314/commclinicsinitmeddirectorinfomgmtassmntsurv_pdf. [Accessed 24 March 2012].

[43] J. DiNardo, "Natural Experiments and Quasi-natural Experiments," in *The New Palgrave Dictionary of Economics, Second Edition*, Palgrave Macmillan, 2008.

[45] Wolf.